The power of meta-analysis: a challenge for evidence-based medicine

The power of meta-analysis: a challenge for evidence-based medicine

ABSTRACT

This paper discusses the outstanding problem of replicability of empirical data in the context of recent work on meta-analysis, especially within the field of evidence-based medicine. Specifically, it deals with the methodological issue of how to determine the degrees of heterogeneity between different collected studies. After critically reviewing the standard measures used to quantify meta-analytical heterogeneity, we argue that they should be revised in such a way to take into account the statistical power of the individual studies. We thus propose some new measures of heterogeneity. Subsequently, we apply them to re-assess concrete case-studies from clinical research, thereby showing explicitly how the relevant values of heterogeneity diverge from those obtained with the original measures.

Keywords: meta-analysis; heterogeneity; power; evidence-based medicine

1 Introduction

The possibility of reproducing the results of an experiment is one of the cornerstone of the modern scientific method. For one, the fact that repeating an experiment under fixed conditions yields the same outcome over and over again is regarded as a sign that such an outcome is reliable, provided that the experimental procedure is adequately carried out. Moreover, it enables one to communicate the results of one's research to the scientific community thanks to the fact that different scientists can reproduce the same experiment independently, thereby enhancing objectivity, or at least inter-subjectivity, of the experimental findings. In light of this, Popper [14] went as far as characterizing reproducibility as the "hallmark of science", in that he claimed that single occurrences that cannot be reproduced should have no place in science. Yet, reproducibility proves too strong a requirement, since one can

seldom obtain the exact same outcomes in different experiments. So, scientific practice often resorts to the concept of replicability, which only requires that, if the same experimental procedure is repeated or carried out independently by different experimenters, the results ought to be similar rather than identical. Arguably, the notion of similarity is somewhat vague, yet replicability seems well-suited to those fields of science, such as the social sciences and medicine, in which the experimental results have remarkable statistical features and one needs to amalgamate evidence from different studies.

Nevertheless, due to the increasing availability of large sets of data from diverse sources, in the last decade even this weaker requirement has been called into question, thereby contributing to what has been named the *replicability crisis*. The predicament is perceived as being particularly dramatic in the field of evidence-based medicine, where one employs experimental procedures like randomized controlled trials that rely on statistics in order to limit the impact of biases and to evaluate the potential efficacy of a treatment. Arguably, the problem of replicability arises inasmuch as the results of different studies prove to be highly heterogeneous, and hence it becomes difficult to determine the extent to which they can be regarded as similar. In clinical research, *meta-analysis* is a widely used statistical method that aims to combine the results of multiple collected studies, so as to establish the extent to which there is agreement among them. So, in the context of meta-analysis, the problem of replicability requires one to quantify and interpret the relevant degrees of heterogeneity between the collected results. Here, we focus on the question what is the proper measure of heterogeneity to apply to evidence-based medicine, and we propose a modified method to quantify heterogeneity that revises the standard measures.

The paper is organized as follows. In section 2, after explaining the sense in which meta-analytical heterogeneity is related to the problem of replicability, we review the two main measures of heterogeneity adopted in the literature (section 2.1), namely the Q index and the I^2 index, by emphasizing their limitations. In the following section 3, we proceed to argue that, in order to properly quantify heterogeneity, one ought to take into account the conditional probability that a true effect be correctly detected, or that a false null hypothesis be correctly rejected. That leads us to develop a modified measure based on the concept of a posteriori statistical power (section 3.1), which properly readjusts the relative weight of each individual study. We label such a measure I^2_r so as to emphasize that it *r*evises the I^2 index. Section 4 is devoted to apply our newly defined measure of heterogeneity to specific examples of clinical research, where we compare it with the other standard heterogeneity indexes. Finally, in section 5, we explain in what sense including statistical power in the assessment of heterogeneity helps us

better understand the problem of replicability in meta-analysis.

2 Meta-analysis in the replicability crisis

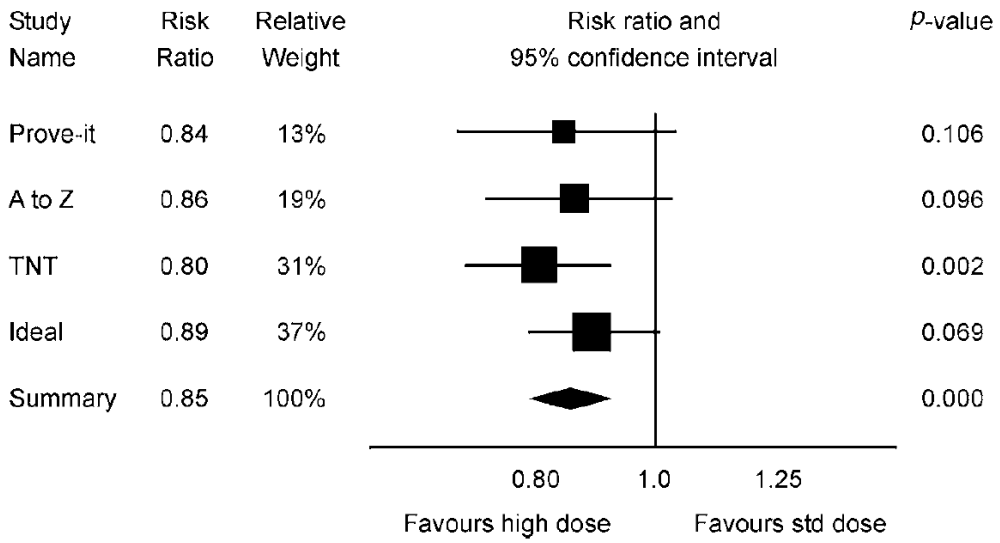
Meta-analysis is a procedure by which the results of multiple studies are collected and combined. It has a long tradition in statistics tracing back to Simpson and Pearson [19] and, even though it has been mainly used in social sciences in the 1970s, nowadays it is extensively applied in clinical research following the evidence-based approach to medicine [1]. Indeed, meta-analysis is regarded as a powerful method to provide the highest level of statistical evidence. In the contest of clinical research it fulfills a variety of objectives: most notably, providing a synthesis of the results of individual studies; investigating the heterogeneity in the results and differences in methods among studies; overcoming small samples sizes of individual study to detect effects of interest and investigate endpoints requiring larger sample sizes; increasing precision in assessing effects in subsets of patients; determining if new studies are needed to further investigate the issue; generating new clinical hypotheses [25]. The use of meta-analysis has become so widespread in evidence-based medicine that, as Stegenga (2009) [21] observes, it is commonly taken to provide the "platinum standard of evidence". Variation between studies included in the collected sample, namely heterogeneity, is one of the key components upon which the meta-analysis is reviewed. The underlying idea is that, if there is a lot of heterogeneity, the results of the meta-analysis cannot be considered as a reliable and generalizable estimate of the pooled effect size. If so, though, the problem of replicability of the experimental results presents itself in evidence-based medicine, as well as in any other field adopting meta-analysis as a statistical method to combine multiple studies.

When it comes to evaluating evidence and combining the results of different studies, there arise plenty of philosophical issues. Just to echo Worrall (2007), "we need guidance in particular on what these different types of evidence are to be amalgamated and in particular on what to do when different types of evidence seem *prima facie* to clash" [27]. Meta-analysis covers a key role in evidence-based medicine for the purpose of amalgamating evidence from different sources in face of heterogeneity. That said, in philosophy of science meta-analysis has not yet received the same attention as in the statistical sciences. To our knowledge, in the philosophical literature one of the few systematic discussions of meta-analysis in the context of evidence-based medicine is given by Stegenga [21], [22], and his diagnosis is rather critical[1]: according to him, meta-analysis should not be regarded as providing

[1]See also [6] for another relevant discussion present in philosophical literature.

the platinum standard of evidence because the very process of collecting the relevant data to which the statistical methods are applied is already beset with subjective factors and intrinsic biases . Here, however, we are concerned with a different issue than subjectivity. It is the methodological issue of how to evaluate and quantify heterogeneity. It is connected with the problem of replicability in the sense that the extent to which the results of different studies in a sample can be regarded as similar (one may say "robust" across the sample), as required by the notion of replicability, effectively depends on how heterogeneous the studies included in the meta-analysis are. Thus, in order to cope with this problem, one needs to understand what factors are responsible for the diversity between the results, as well as to determine the best mathematical measure for the degrees of heterogeneity.

To see how meta-analysis works, let us begin with the informal presentation by Stegenga [21], and then illustrate it with an example from evidence-based medicine. According to it, the general methodology employed in meta-analysis comprises four subsequent steps. That is, (i) selecting which primary studies are to be included; (ii) calculating the magnitude of the effect one wishes to investigate; (iii) assigning a weight to each study; and (iv) calculating a weighted average of the effect magnitudes. Step (i) is of course the starting point in meta-analysis, since the aim is to compare a fix sample of collected studies whose results are then combined. As we shall explain in greater detail below, a big source of heterogeneity comes just from the selection of very diverse studies. The second step (ii) is more formal: once the relevant effect E has been identified, one must determine the value of its magnitude, possibly under certain statistical conditions. To fix the terminology, in the example of clinical trials the effect size corresponds to a numerical value that reflects the magnitude of the treatment effect. Mathematically, that expresses the strength of a relationship between two variables: for instance, the risk ratio of the reduction of cardiovascular outcomes with high dose versus standard dose of statins. When the risk ratio (the ratio of the incidence in the two groups of patients) is equal to 1, it means that there is no statistical difference between the two treatments. Instead, if a risk ratio is less than 1 it means that the risk is lower in the high-dose group, whereas a risk ratio greater than 1 means that the risk was lower in the standard-dose group. In the following figure, taken from [2],



solid squares are used to depict how the size of each study varies. The p-value encodes the probability of obtaining a test statistics being at least as extreme as the one actually observed, on the assumption that the null hypothesis H_0 is true. Then, step (iii) prescribes that, under the same operating statistical conditions, we also assign a weight to each study. This is actually a crucial point for the assessment of heterogeneity, to which we shall focus in section 3. For now, it is worth just noting that the way in which weights are assigned depends on our hypotheses about the distribution of effect sizes from which the studies were sampled[2]. Finally, as prescribed in step (iv), the summary effect is calculated as the weighted mean of the individual effects, which are represented in the figure by a diamond. In this specific case, the summary risk ratio is 0.85, indicating that the risk of cardiovascular problems is lower for patients receiving the high dose rather than for patients receiving the standard dose. This gives a rough illustration of the general methodology followed in a meta-analysis.

Sources of heterogeneity arise throughout the process. As Rücker et al. pointed out [16], one can identify three sources of heterogeneity plaguing meta-analysis in the context of evidence-based medicine. The first one is clinical baseline heterogeneity, which is due to the differences among sample characteristics between the studies: for instance, various patients may belong to rather different age groups. The second source is traced back to the

[2]In particular, under the fixed-effect model, one assumes that all studies in the analysis have the same true effect size, and the summary effect is our estimate of this common effect size; under the random-effects model, instead, one assumes that the true effect size varies from study to study, and the summary effect is our estimate of the mean of the distribution of effect sizes (see [2])

statistical heterogeneity found in the collected outcomes. In general, if the outcomes vary considerably from study to study, we are not in a position to determine how effective a clinical treatment really is. Moreover, that dilutes the confidence in the pooled effect, too. Thirdly, one can have heterogeneity coming from other sources, like design-related heterogeneity. In terms of the general scheme outlined by Stegenga, it seems that the factors responsible for heterogeneity mostly intervene at the stage in which one sorts out the primary studies, i.e. step (i); and, especially as regards the second statistical source, they also affect the way in which one computes the magnitude of the effect, i.e. step (ii). Arguably, these factors are accountable for the intrinsic subjectivity Stegenga imputes meta-analysis for. However, differently from Stegenga, the relevant issue we are concerned with is a methodological one: that is, granted that there is already heterogeneity, as it does arise from the above listed sources, one needs to establish how to properly quantify it. In this respect, our own focus is on the other two steps (iii) and (iv), which effectively shape the particular method one adopts to compute the degrees of heterogeneity. Indeed, we contend that the appropriate measure of heterogeneity across multiple studies strongly depends on how exactly one assigns the weights to the individual studies in step (iii), since that determines the overall value of heterogeneity calculated through the subsequent step (iv) as an average across the studies. In order to see why that is the case, here below we present and discuss the standard measures of heterogeneity, so as to emphasize how they explicitly depend on the relative weight of each study. The quantitative method we will subsequently develop aims to contribute to limit the impact of the statistical component of heterogeneity, so that the remaining amount can be just traced back to design-based factors and, most importantly, to clinical factors.

2.1 Quantifying *Heterogeneity*

One of the standard methods to quantify heterogeneity is provided by the Q index, which was first proposed by Cochrane in 1954 [3]. Formally, for an effect detected in the studies included in a meta-analysis, given the number k of collected studies, namely the *sample size*, such a measure is defined as follows:

$$Q = \sum_{i=1}^{k} w_i (T_i - \bar{T})^2 \tag{1}$$

where T_i is the *effect size* of each individual study i with weight w_i and $\bar{T}$ is the mean effect size of all the studies. In other words, Q is constructed as the

weighted sum over i of the squared differences between each individual study effect and the pooled effect across studies. The contribution of each study i is thus given by two components. One is the quantity $(T_i - \bar{T})^2$, which measures the deviation of the size effect T_i of the study with respect to the mean $\bar{T}$ of the sample of size k. It tells us how much study i diverges from the other collected studies. The other component is the relative weight w_i. In particular, in the simple case of a fixed-effects statistical model, wherein all the studies in the population are supposed to be conducted under the same conditions so that their summary statistics result from estimates with common mean, it is standard to compute the weight of each study as the inverse-variance, that is $w_i = \frac{1}{s_i^2}$ with s_i being the deviation of study i from the mean. Since there are different ways to assign weights to the individual studies, though, it is clear that how exactly one decides to do so affects the overall value of heterogeneity. Indeed, studies for which the effect size T_i largely deviates from the mean size $\bar{T}$ will contribute more than if they have high weight rather than low weight, thereby increasing heterogeneity. We will present a different method to assign weights in the next section. What we wish to stress here is that the value of the Cochrane index is known to increase when the sample size k grows. Unfortunately, though, it is also known that, if the number k of studies remains low, the ability of Q to detect heterogeneity proves rather poor. As a consequence, one can hardly take it as a generally reliable measure [17].

In order to improve on the Cochrane index, Higgins et al. [10] introduced another measure of heterogeneity, which is labeled the I^2 index. It is defined by the following formula:

$$I^2 = (\frac{Q - df}{Q}) \times 100\% \tag{2}$$

where Q is the Cochrane index and the term $df = k - 1$ is given by the number of studies minus one degree of freedom. Technically, the I^2 statistics represents the percentage of variability in the effect sizes that is not caused by the sampling error. As such, it corresponds to the proportion of actual between-study heterogeneity with respect to the total amount of heterogeneity in the meta-analysis, and thus it expresses the mutual inconsistencies of the various studies' outcomes. In principle, according to the above formula (2), negative values could be possible if $Q < df$, but Higgins et al. suggest that such cases should be interpreted in the same way as $I^2 = 0\%$. A zero value means that there is no actual between-study heterogeneity, whereas $I^2 = 100\%$ indicates that all heterogeneity is due to between-study heterogeneity. Typically, for $Q > df$ one obtains values between these two extreme cases. It is important to stress that the I^2 statistics involves some degrees of

uncertainty: therefore, it is recommended to present it together with a confidence interval (CI), especially when the number of studies k is low. This measure of heterogeneity has some advantages with respect to the Cochrane's Q index in that, unlike the latter, the I^2 index is comparable between different meta-analyses. Indeed, the I^2 statistics is not at all sensitive to changes in the sample size. Nevertheless, there are also limitations with the use of such a standard measure. For, it suffers from some arbitrariness as regards how exactly the percentages should be interpreted: for instance, Higgins et al. [9] stipulate that the heterogeneity is mild if I^2 is less than 30% and notable if it is greater than 50%, whereas in [10] heterogeneity is low when $I^2 = 25\%$, moderate when $I^2 = 50\%$ and high when $I^2 = 75\%$. Furthermore, in the presence of limited data, the uncertainty on the value of I^2 is very large, in particular for the common situation of meta-analyses with two or three studies [11].

Therefore, while the I^2 index improves on the Q index, it is doubtful that, as it stands, it can serve as a proper measure of heterogeneity in many important cases. To add to the above limitations[3], we wish to observe that the definition of I^2 given in equation (2) is sensitive to the choice of the weights w_i of each individual study appearing in equation (1) for Cochrane's Q index. However, there are different possible ways to assign such weights in the definition of heterogeneity, and the standard choice employed in formulas (1) and (2) may not be adequate. In the next section we formulate a modified method to quantify heterogeneity in meta-analysis, which revises both the Q and the I^2 measure, in the sense that the choice of the weights hinges on another important aspect of meta-analysis, namely the *statistical power* of the individual studies included in the sample. In fact, we argue that failure to incorporate statistical power is yet another weakness of the standard indexes, over and above the already well-known limitations.

3 How much statistical power is needed?

Even though the heterogeneity of the included studies is important when evaluating the meta-analysis, it is just one of the aspects to be considered. Taking into account the statistical power of the original studies provides a different point of view for reviewing the meta-analysis. The guiding idea is that studies with higher statistical power should have a stronger impact on the conclusions of the meta-analysis than those with lower statistical power.

[3]It should be stressed, though, that arbitrariness and uncertainty, as well as poor effectiveness for small sample size, are of course limitations that are common to other measures, too.

We thus submit that for the purpose of quantifying heterogeneity one ought to include the statistical power of the studies under consideration. Accordingly, here below in section 3.1 we construct a new measure of heterogeneity that properly revises the standard I^2 index introduced in section 2.1. Before doing so, we explain in great details what is the role of statistical power when assessing single studies and why it must be incorporated in meta-analytical heterogeneity.

To introduce the concept of power of a study, let us explain first of all that the types of error one ought to avoid in meta-analysis are the same as in statistical test theory. For, suppose the null hypothesis H_0 states that there is no effect across the selected individual studies: then, if a certain effect E is detected in a study, H_0 proves false and hence it should be rejected. Type I errors occur when one rejects the null hypothesis even if H_0 is true. Correspondingly, in a meta-analysis one would commit to a false positive if by combining the results of the collected studies one detects an effect E even though the effect is actually false. The probability of a Type I error is denoted by α. Type II errors, instead, occur when one fails to reject the null hypothesis even if H_0 is false. Correspondingly, in a meta-analysis one would commit to a false negative if by combining the results of the collected studies one does not detect an effect E even though the effect is actually true. The probability of a Type II error is denoted by β. Type I errors are typically considered to be about four times more dangerous than Type II errors. To minimize risk, it is thus customary to first set the minimum acceptable probability α to the conventional value 0.05, meaning that the likelihood of rejecting a true null hypothesis or to detect a false effect cannot exceed 5%. The one-to-four trade-off between α and β then gives a minimum acceptable probability of Type II errors being four times bigger than that of Type I errors, that is $0.2 = 20\%$. Of course, what matters most in a meta-analysis is not just to avoid Type I and Type II errors but also to reveal true negatives and true positives, so as to determine whether a study correctly detects a true effect and correctly rejects a false null hypothesis and to estimate with better precision the true effect magnitude or effect size.

The *statistical power* of a test of statistical significance encodes the probability that the test correctly rejects the null hypothesis when H_0 is false. Correspondingly, it quantifies the likelihood that an effect E is detected across the studies included in a meta-analysis when E is actually present. In fact, the statistical power π of each individual study is the (conditional) probability that the study reveals a true positive. As such, it is inversely related to the likelihood of committing a Type II error, and hence it is given by $\pi = 1 - \beta$. It follows that increasing statistical power diminishes the probability of false negatives. Underpowered studies are thus less reliable than

studies with high power. For this reason, it is essential to take statistical power into account when carrying out a meta-analysis. Note that the statistical power depends upon the magnitude of the effect to be detected as well as the sample size: so, the larger the effect size the higher the statistical power at the same sample size. In particular, given the one-to-four trade-off between $\alpha = 0.05$ and $\beta = 0.2$, the minimum acceptable power for a study is commonly recommended to be 80%, that is $1 - 0.2 = 0.8$. Surely, we do not advocate individual statistical power as a stand-alone criterion for a study to be included in a meta-analysis. Indeed, according to, e.g., the review by Turner and colleagues [23], most meta-analyses include studies that do not have enough statistical power to detect a meaningful (or clinically significant in biomedical research) effect. However, it is also true that failure to rely on power has an impact on the overall conclusions of a meta-analysis, especially when (i) the conclusions are drawn from the results of underpowered studies, or when (ii) the difference in statistical power of single studies is not taken into account. Hence, an evaluation of the power of each studies ought to be included in the estimation of heterogeneity among the collected studies. In light of these remarks, it thus seems quite natural to demand that a proper measure of meta-analytical heterogeneity should include the values of statistical power of the individual studies in the selected sample.

For purposes of computing the value π_i for each study i, recall that statistical power is defined as a conditional probability. Here, it is important to distinguish between retrospective and a posteriori power [28]. On the one hand, retrospective statistical power is the conditional probability of correctly rejecting a false null hypothesis after H_0 has been rejected, that is $Pr(H_0\ false \mid reject\ H_0)$. As such, it is computed after a study has been conducted and one has already decided whether to reject the null hypothesis or not (and in fact, it is also known as post-study probability). On the other hand, a posteriori statistical power is the conditional probability of correctly rejecting a null hypothesis when H_0 is actually false, that is $Pr(reject\ H_0 |\ H_0\ false)$. As such, it is computed before a study is conducted, so as to estimate the sample size necessary to detect a meaningful effect. In general, these two measures of power yield different values, except in the trivial case in which the prior probability $Pr(H_0\ false)$ that H_0 be false and the prior probability $Pr(reject\ H_0)$ of rejecting H_0 are equal. Thus, when computing π_i for each study i, one must choose whether to adopt retrospective or a posteriori statistical power. Fallacies due to retrospective statistical power are actually typical and, in a certain sense, its use is demonstrated to be fundamentally flawed (see, for instance, [26] for a critique). So, a posteriori power seems to be a more suitable choice in order to embed statistical power into the measure of heterogeneity, and indeed it is taken to be

standard notion by practicing statisticians.

This fact becomes transparent if we cast the above definitions in terms of whether a certain effect E is detected or not. Retrospective power quantifies how likely it is that, if the effect E has been detected in study i, such an effect is actually true. It thus tells us the extent to which we are licensed to infer that the effect E is real from the fact that it has been detected. Instead, a posteriori statistical power quantifies how likely it is that, if the effect E is actually true, such an effect is detected by study i. Here, differently from the retrospective case, the power is prospective in that one presupposes that the effect E is real: one then determines what is the probability that it be detected in any study i in the sample. A posteriori statistical power thus corresponds exactly to the probability of obtaining a true positive. In fact, it is usually computed before conducting a study, in such a way to estimate the sample size necessary to detect a meaningful effect at the level of type I error. To put the distinction more technically, retrospective statistical power is a conditional probability with respect to the size effect detected by the study, that is one fixes the size effect and lets the sample size vary; a posteriori statistical power, instead, is a conditional probability with respect to the actual sample size k, that is one fixes the sample size and lets the size effect vary. Since we adopt the latter notion of statistical power, for the sake of evaluating the power values π_i of the studies collected in a sample of fixed size $i = 1, ..., k$ we suggest to consider three distinct scenarios, involving the detection of *(B)* a small effect size, *(C)* a medium effect size and *(D)* a large effect size, as classified by Cohen [4][4].

We are now in position to introduce a new measure of heterogeneity that takes into account the statistical powers of the individual studies. Specifically, we propose to modify the standard heterogeneity indexes Q and I^2 so as to incorporate the a posteriori statistical power of each study entering into a meta-analysis. In our view, this enables one to provide a more refined judgement of meta-analytical heterogeneity.

3.1 A revised measure of heterogeneity

The mathematical formulation of our proposed measure hinges on a revision of the I^2 index, whereby the weights of the studies appearing in the Cochrane index Q in formula (1) are recomputed based on statistical power, namely the probability of correctly detecting a true effect, and the conventional minimum acceptable power π_0, that is $0.8 = 80\%$. Accordingly, if π_i denotes the a

[4]Let us note that Ioannidis [12] pointed out that effect sizes of newly discovered true associations are essentially inflated on average if power is not considered. Our approach thus tries to deal with the issue of the magnitude of effect size based on statistical power.

posteriori statistical power of the individual study i, the adjusted weight becomes

$$\tilde{w}_i = \frac{\pi_i}{\pi_0} w_i \tag{3}$$

where w_i is the original weight of study i. The readjusting factor depends on whether the statistical power π_i of study i is greater, equal or smaller than the minimal accepted power π_0, so that the adjusted weight $\tilde{w}_i$ becomes greater, equal or smaller than the original weight w_i, respectively[5]. In other words, studies with high statistical power increase their relative weight, whereas under-powered studies are assigned smaller weights.

By replacing the weights $\tilde{w}_i$ expressed by (3) for the original weights w_i in formula (1) of the Cochrane's index, one obtains the new index

$$Q_r = \sum_{i=1}^{k} \tilde{w}_i (T_i - \bar{T})^2 \tag{4}$$

with T_i being the effect size detected by each study and $\bar{T}$ the mean effect size. This means that, for fixed deviations $(T_i - \bar{T})^2$, studies with high statistical power contribute more than under-powered studies to the value of heterogeneity. So, compared with the standard measure, when the detected effect size T_i largely deviates from the mean effect size $\bar{T}$, studies with low π_i have less impact on the heterogeneity of the sample since the adjusted weight $\tilde{w}_i$ is less than the original weight w_i, whereas studies with high π_i have a greater impact since the adjusted weight $\tilde{w}_i$ is greater than the original weight w_i. The revised Q_r index have the same value of Cochrane's Q index if the power of all the collected studies is exactly equal to the minimal acceptable power π_0.

Similarly to what was done in section 2.1, the next step is to define the proper measure of heterogeneity as a straightforward revision of formula (2) for the I^2 index based on the new Q_r index, as follows:

$$I_r^2 = \left(\frac{Q_r - df}{Q_r}\right) \times 100\% \tag{5}$$

where again df represents the total number of collected studies k minus one degree of freedom, and hence it depends on the sample size. This index varies depending on the value of the new Cochrane index calculated with the adjusted weights $\bar{w}_i$: so, the higher Q_r with respect to df, the higher the I_r^2

[5]Let us note that, contrary to the original weights w_i, summing over the new weights $\tilde{w}_i$ of all k studies does not necessarily yield a value equal to 1, just owing to the presence of the readjusting factor $\frac{\pi_i}{\pi_0}$.

index. In particular, the I_r^2 index vanishes if Q_r is equal to df, meaning that there is no heterogeneity in the sample. In addition, as it is customary, we set $I_r^2 = 0\%$ also whenever $Q_r < df$. Of course, our measure of heterogeneity (5) is also sensitive to the individual statistical powers π_i of the $i = 1, ..., k$ studies included in the sample, which appear in equation (4) for the Q_r index: therefore, even according to the I_r^2 index, studies with high statistical power have a greater impact on the meta-analytical heterogeneity of the sample than underpowered studies.

In general, the indexes Q_r and I_r^2 we have thus constructed yield values that are quite different from the standard indexes Q and I^2. Such values are more informative about the heterogeneity across a sample of $i = 1, ..., k$ collected studies, because they take into account also the statistical power π_i of each individual study. Using the adjusted weights $\tilde{w}_i$ assures that, for an effect E, studies with high statistical power increase meta-analytical heterogeneity more than underpowered studies, when the effect size T_i deviates from the mean effect size $\bar{T}$. Armed with our refined measure of heterogeneity, we can now proceed to re-assess the standard meta-analyses of concrete examples of clinical research with a method that incorporates statistical power, in the hope to better cope with the problems of evidence amalgamation and replicability arising in evidence-based medicine.

4 Case-studies from evidence-based medicine

In the present section we apply our proposed measure of heterogeneity I_r^2 to two real-life examples of meta-analyses, so as to show how explicit calculations based on formula (5) yield different results than the standard measure I^2. To do this we re-run the meta-analysis of Crins et al. [5], which examines the effect of Interleukin-2 receptor antibodies (IL-2RA) on safe immunosuppression after liver transplantation in children. In this work, which adopts the standard Q and I^2 indexes, the efficacy of the therapy was evaluated on different outcomes: specifically, the reduction of *(i)* steroids-resistant rejections (SSRs) and *(ii)* acute rejections (ARs)[6]. We choose to work with these case-studies because they highlight some of the critical issues of meta-analysis we previously discussed, like the reduced sample size (namely the small number k of included studies), which tend to affect heterogeneity.

We proceed by discussing the two case-studies *(i)* SSRs and *(ii)* ARs in the order. For both, we first present the results obtained by the meta-analysis of Crins et al., in particular the values of the original Cochrane's Q index and

[6]For completeness, let us mention that Crins et al. also run their meta-analysis for the cases of graft-loss and death, which we do not address here.

the original I^2 index, as given by equations (1) and (2), respectively. Then, we determine the statistical power of each study i. Given that our method uses a posteriori statistical power, the effect size varies while the sample size remains fixed: to this extent, we compute the a posteriori statistical power π_i of each study in three distinct scenarios, involving the detection of *(B)* a small effect size, *(C)* a medium effect size and *(D)* a large effect size, as classified by Cohen [4]. Finally, on the basis of the weights $\tilde{w}_i$ adjusted in accordance with formula (3), we calculate the revised Q_r index and the I_r^2 index by means of equations (4) and (5), respectively. The results of our meta-analysis in the three scenarios can then be compared with the original results of Crins et al. for both cases *(i)* SSRs and *(ii)* ARs. Our calculations have been performed by using *R* version 3.5.0 as in [15] and metaphor package [24].

Let us begin with case *(i)* concerning steroids-resistant rejections. Here, the sample size is relatively low, since the meta-analysis of Crins et al. includes just $k = 3$ different clinical studies. In order to estimate the between-study variance s_i^2 of each study, which is needed to compute the weights w_i, they employed the DerSimonian–Laird (DL) [7] approach. The results of their meta-analysis on the controlled trials assessing SRRs in paediatric patients with IL-2RA as monotherapy or in combination in liver transplant recipients compared with placebo or no add-on are reported in Figure 1A. Notice that, based on original weights, statistical significance was achieved. Moreover, Cochran's Q index was equal to 2.37, and therefore by means of formula (2) one obtains $I^2 = \frac{2.37-2}{2.37} = 15.59\%$. So, in this case heterogeneity was not particularly high, although it was given with a confidence interval $95\%CI : (0; 60)\%$, which ranges just from very low to moderate[7]. Let us now compare these results with those obtained with our proposed method under three distinct scenarios, involving the detection of *(B)* a small effect size, *(C)* a medium effect size and *(D)* a large effect size. The respective values of the a posteriori statistical power π_i we calculated for each study are shown in Table 1, together with the overall values of the revised heterogeneity index I_r^2. The new meta-analysis derived by using the weights adjusted according to our method is shown in Figure 1. It can be observed that in the scenarios of small and medium effect, the revised meta-analytic risk-ratio summary estimate shows a smaller risk reduction and the statistical significance is no more reached, since the null value 1 is inside the 95%CI. What is more, heterogeneity completely vanishes, that is $I_r^2 = 0\%$, in scenario *(B)* of a small effect size and scenario *(C)* of a medium effect size for which the adjusted

[7]Note that in this case the source of heterogeneity mainly arises from differences in the design of the studies themselves (for instance, only the study by Heffron [8] is randomized).

Table 1: A posteriori power computed on the basis of the real sample size

	Power		
	Small Effect Scenario	Medium Effect Scenario	Large Effect Scenario
Heffron 2003	12.1%	49.2%	87.4%
Granschow 2005	18.0%	73.8%	98.6%
Gros 2008	14.7%	61.4%	94.9%
I_r^2 (95%CI)	0% (0; 1)%	0% (0; 89)%	27% (15; 73)%

Cochrane index yields values $Q_r = 0.44$ and $Q_r = 1.79$, respectively. That is due to the fact that all the three studies have relatively low statistical power, and hence their contribution to heterogeneity is not very significant, to a point that Q_r is less than df, which is equal to 2 for a sample including three studies. Instead, in scenario *(D)* of large effect size the value of heterogeneity has slightly raised compared with Crins et al., in that we obtained $Q_r = 2.76$ and $I_r^2 = 27\%$ for the adjusted indexes. Here, the discrepancy with respect to the original values of Q and I^2 is explained by the fact that the statistical powers π_i of all studies have increased above the minimal acceptable power π_0, and therefore the contributions coming from their adjusted weights $\tilde{w}_i$ are higher than from their original weights w_i.

Let us now turn to the second case-study *(ii)* concerning acute rejections. The meta-analysis considered by Crins et al. includes six controlled trials assessing the effect of IL-2RA on ARs, and hence we have a larger sample size than in the previous case-study *(i)* concerning steroids-resistant rejections. Figure 2A shows the original results published in [5]. The original Q for Cochrane statistics was equal to 14.51, statistically significant (p = 0.012), and therefore it follows from equation (2) that $I^2 = 66\%$. Heterogeneity is higher than in case *(i)* because the sample size has increased. Similarly to our treatment of case-study *(i)*, we computed the a posteriori power under the three scenarios of detecting *(B)* a small, *(C)* a medium and *(D)* a large effect. The results we obtained for each study are reported in Table 2, along with the values for our revised measure of heterogeneity I_r^2. The meta-analyses under the three scenarios obtained by applying the adjusted weights $\tilde{w}_i$ described by formula (3) are depicted in Figure 2B, Figure 2C, Figure 2D. Let us stress that the revised heterogeneity index I_r^2 vanishes for scenario *(B)* of small effect size for which the adjusted Cochrane index is $Q_r = 2.23$. This is due to the fact that such a value is smaller than df, which here is equal to 5. Moreover, the index I_r^2 is not zero, but still lower than the original value obtained by Crins et al. in scenario *(C)* of medium size effect since by equation (5) we obtained $I_r^2 = 44\%$ from the adjusted Cochrane index

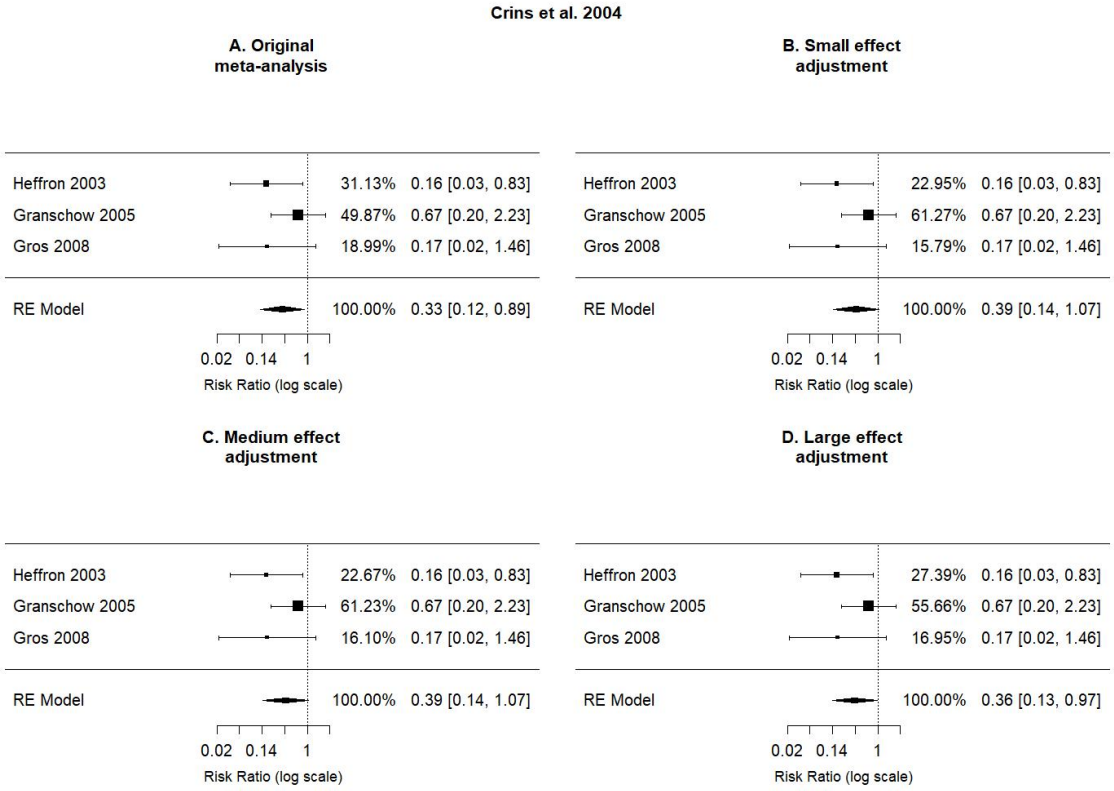


Figure 1: Treatment effect on SRRs estimated in the original meta-analysis (A), adjusting the weights by the statistical a posteriori power under the scenario of a small effect size (B), a medium effect size (C) and a large effect size (D).

$Q_r = 8.95$. In fact, the latter is greater than df, so that there must be some heterogeneity present, yet it is smaller than the original Cochrane's index Q. A smaller value of heterogeneity indicates that for some of the studies with higher statistical power, and hence with high adjusted weights $\tilde{w}_i$, the detected effect size T_i does not diverge much from the mean effect size $\bar{T}$. Finally, the index I_r^2 remains roughly the same in scenario *(D)* of large size effect, where we obtained $I_r^2 = 68\%$ from the adjusted Cochrane index being $Q_r = 15.51$. Note that, except for the relatively underpowered study Schuller [18], all the other studies have statistical power above the minimal acceptance power π_0. Beside the remarks already made, we just add that the behaviour of the revised degrees of heterogeneity across the three scenarios shows that the index tends to increase as the size effect grows. In fact, also the individual statistical power π_i of each study becomes progressively higher

Table 2: A posteriori power computed on the basis of the real sample size and readjusted I_r^2 statistics

Studies	Power		
	Small Effect Scenario	Medium Effect Scenario	Large Effect Scenario
Heffron 2003	12.1%	49.2%	87.4%
Gibelli 2004	11.6%	46.5%	84.9%
Granschow 2005	18.0%	73.8%	98.6%
Schuller 2005	8.4%	26.9%	57.4%
Spada 2006	13.6%	56.4%	92.4%
Gros 2008	14.7%	61.4%	94.9%
I_r^2	0% (0; 62)%	44% (0; 78)%	68% (24; 86)%

when the size effect grows from being small to medium and then to large, thereby increasing the adjusted weight $\tilde{w}_i$.

5 What statistical power can tell us about replicability in meta-analysis

The above comparison between the original meta-analysis by Chris et al. and the approach we advocate well illustrates how the values of heterogeneity diverge when being computed with the adjusted indexes Q_r and I_r^2 as opposed to the standard indexes Q and I^2. That is just one possible example taken from clinical research, yet our method can be applied virtually to any meta-analysis even beside the context of evidence-based medicine, provided that one can properly calculate the statistical power of the studies included in the sample. In fact, we submit that our suggested way to quantify heterogeneity being conceptually motivated by the importance of incorporating statistical power is not only relevant to the philosophical foundations of statistics, but it also provides a valuable tool in the assessment of meta-analysis to practicing statisticians dealing with those empirical sciences in which evidence from various sources needs to be combined. Here, we wish to conclude our discussion by returning to the replicability crisis, so as to explain the sense in which our proposal can contribute to cope with it.

Recall that the problem of replicability in evidence-based medicine rests on the fact that the results of different studies about the same clinical effect fail to be similar. Allegedly, a high value of meta-analytical heterogeneity indicates that the sample results are very diverse, and therefore they could

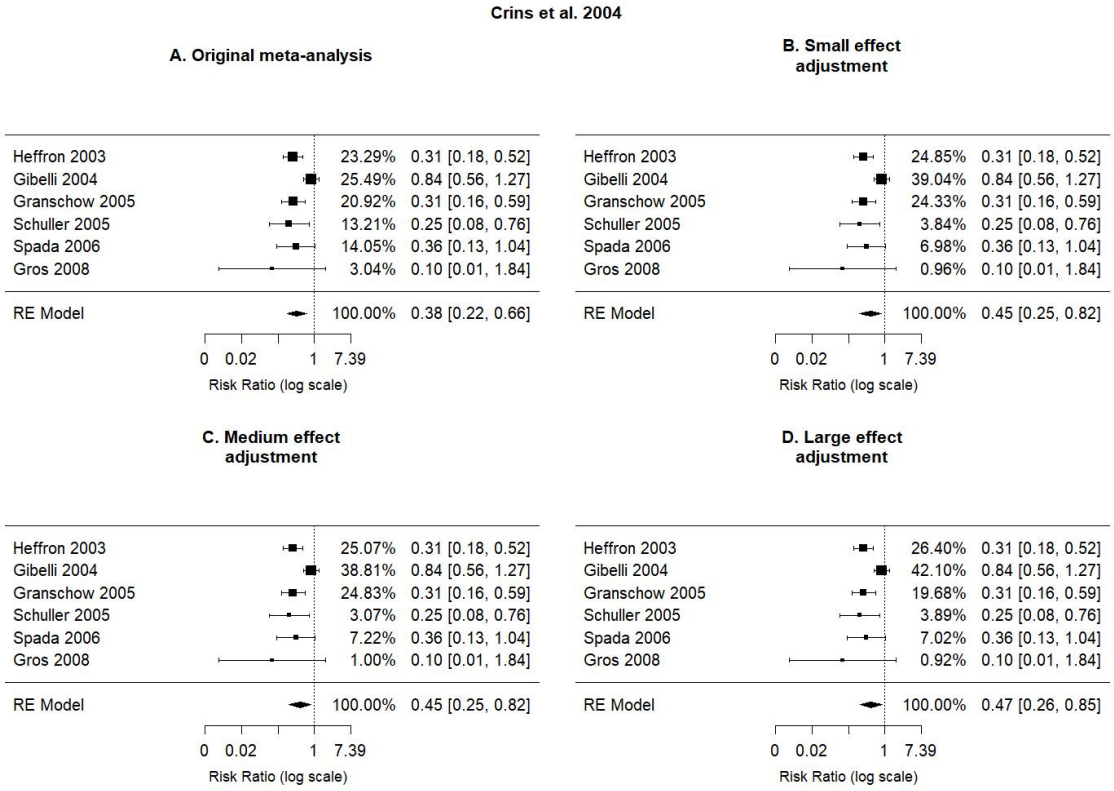


Figure 2: Treatment effect on ARs estimated in the original meta-analysis (A), adjusting the weights by the statistical a posteriori power under the scenario of a small effect size (B), a medium effect size (C) and a large effect size (D).

hardly be regarded as similar. The methodological issue how to quantify heterogeneity thus becomes crucial in order to determine the extent to which the effect under investigation can be said to be reliably detected across the sample. In principle, the problem of replicability is resolved, or at least assuaged, by reducing the degrees of heterogeneity across the sample. In this respect, however, it should be made clear that the purpose is not to find a measure of heterogeneity that just yields smaller values than the standard indexes. Else, one could simply construct a programmatic measure having the same mathematical form as (1) in such a way to assign conveniently low weights w_i's to those studies whose effect size T_i largely deviates from the mean effect size $\bar{T}$, thereby decreasing their contribution to the overall weighted average. But that would of course be an *ad hoc* solution of the problem of replicability. Instead, the purpose is to find a proper measure of

heterogeneity grounded on relevant and conceptually sound considerations, like the need of incorporating the statistical power of each individual study as in our method, which can tell us in a reliable way whether or not the results are (sufficiently) similar across the sample. So, rather than providing a definite solution, our proposal aims at establishing some of the conditions under which the problem of replicability would actually present itself or not.

For instance, in the concrete example from evidence-based medicine we previously discussed, if the effect size is not large, our revised indexes yield lower values of heterogeneity than the standard indexes in the original meta-analysis by Crins et al. both in case (i) and case (ii). In fact, I_r^2 even vanishes for scenario *(B)*. This is certainly a remarkable fact. Yet, even so, we do not go as far as claiming that it implies that we resolved the problem of replicability. Indeed, for the large size scenario *(D)* the adjusted values are comparable to the original values, and they actually tend to be slightly bigger. More generally, in a meta-analysis wherein the standard deviation $(T_i - \bar{T})^2$ is rather significant for studies with great statistical power, both Q_r and I_r^2 yield higher values than Q and I^2, respectively, because π_i is greater than the conventional power π_0 and hence the adjusted weights $\tilde{w}_i$ grow with respect to the original weights w_i. Accordingly, there arises the problem of replicability, insofar as the sample exhibits a lot of meta-analytical heterogeneity. By neglecting statistical power, the standard indexes may instead fail to recognize this fact, in that they underestimate the contributions of the most reliable studies included in the sample when the latter detect an effect size that is quite different from the mean effect size. By and large, replicating studies with high statistical power is and should be more difficult than replicating studies with low statistical power. In this regard, the measures of heterogeneity we constructed prove to be a more powerful resource than the standard measures to identify the cases in which the requirement of replicability is called into question in a meta-analysis. In other words, including statistical power in the assessment of the degrees of meta-analytical heterogeneity can tell us in a more effective way than the standard indexes how serious the problem of replicability appears to be across a given selection of studies.

6 Conclusion

In this paper, we have addressed the issue how to define a proper measure of heterogeneity in a meta-analysis of multiple studies designed to detect a certain effect E, which can be applied to concrete examples in evidence-based medicine. This is a methodological issue connected with the problem

of replicability of the experimental results, in so far as the degrees of heterogeneity in a sample of studies indicates the extent to which the results of the latter can be regarded as similar, so as to comply with a basic desideratum for replicability. We have argued that meta-analytical heterogeneity should take into account also the statistical power of the individual studies in the meta-analysis. That has led us to revise the standard measures of heterogeneity, namely the Q and I^2 indexes, in such a way to incorporate the statistical power of each study into their adjusted weights $\tilde{w}_i$, which rescale the original weights w_i by the ratio of the individual statistical power π_i and the minimal statistical power π_0. Our revised measures of heterogeneity have been labelled Q_r and I_r^2 indexes, respectively. According to them, studies with high power contribute more than underpowered studies to the level of meta-analytical heterogeneity, given a certain deviation of the detected effect size with respect to the mean effect size of the sample. We have then applied our proposed method to two published meta-analyses, and we have demonstrated that the values of heterogeneity calculated for the revised indexes Q_r and I_r^2 diverge from the values calculated for the standard indexes Q and I^2.

We thus propose that, in general, all meta-analyses should consider the a posteriori statistical power of individual studies in order to better evaluate heterogeneity. In fact, with respect to the original I^2 index, the revised I_r^2 index strengthens the proportional weights of statistically most reliable studies, thereby improving the overall reliability of the pooled effect size. Moreover, as our examples evidently show, it is crucial to indicate "the effect size level", namely small, medium, or large. For, if the original studies are able to detect only a large effect of a treatment, the pooled effect size cannot be generalized to small and medium effect size scenarios.

References

[1] Bohlin, I. (2012). Formalizing syntheses of medical knowledge: The rise of meta-analysis and systematic reviews. *Perspectives on Science*, 20(3), 273-309.

[2] Borenstein, M., Hedges, L.V., Higgins, J.P., Rothstein, H.R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

[3] Cochrane W.G. The combination of estimates from different experiments (1954). *Biometrics*; 10:101-29. doi:10.2307/3001666.

[4] Cohen, J. Statistical power analysis for the behavioral sciences (1988), Lawrence Earlbaum Associates

[5] Crins N.D., Rover C., Goralczyk A.D., Friede T. (2014). Interleukin-2 receptor antagonists for pediatric liver transplant recipients: a systematic review and meta-analysis of controlled studies. *Pediatric Transplantation* 18: 839–850. DOI:10.1111/petr.12362

[6] Fuller J. (2018), Meta-Research Evidence for Evaluating Therapies. **Philosophy of Science** 85 (5):767-780

[7] DerSimonian R., Laird N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* 7: 177–188. DOI:10.1016/0197-2456(86)90046-2

[8] Heffron T.G. et al. (2003) Pediatric liver transplantation with daclizumab induction therapy. *Transplantation* 2003: 75: 2040–2043.

[9] Higgins, J.P., Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11), 1539-1558.

[10] Higgins J.P.T., Thompson S.G., Deeks J.J., Altman D.G. (2003) Measuring inconsistency in meta-analyses. *BMJ*; 327:557-60. doi:10.1136/bmj.327.7414.557.

[11] Ioannidis J., Patsopoulos N., Evangelou E. (2007) Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* 2007;335:914-6. doi:10.1136/bmj.39343.408449.80.

[12] Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5): 640-648.

[13] Moher D., Cook D.J., Jadad A.R. et al. (1999), Assessing the quality of reports of randomized trials: implications for the conduct of meta-analyses. *Health Technol Assess*;3(12):1-98. doi:10.3310/hta3120.

[14] Popper, K.R (1935). *Logik der Forschung*. Vienna: Julius Springer Verlag.

[15] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

[16] Rücker, G., Schwarzer, G., Carpenter, J. R., Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC medical research methodology*, 8, 79, doi:10.1186/1471-2288-8-79

[17] Schriger, D.L., Altman, D.G., Vetter, J.A., Heafner, T., Moher, D. (2010). Forest plots in reports of systematic reviews: a cross-sectional study reviewing current practice. *International journal of epidemiology*, 39(2), 421-429.

[18] Schuller S., Wiederkehr J.C., Coelho-Lemos I.M., Avilla S.G., Schultz C. (2005), Daclizumab Induction Therapy Associated With Tacrolimus-MMF Has Better Outcome Compared With Tacrolimus-MMF Alone in Pediatric Living Donor Liver Transplantation. *Transplant Proc.*: 37: 1151-1152

[19] Simpson, R.J.S., Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *The British Medical Journal*, 2(2288), 1243-1246.

[20] Soyeon A., Betsy J.B. (2011) Incorporating quality scores in meta-analysis. *J Educ Behav Stat*; 36:555-85. doi:10.3102/1076998610393968.

[21] Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, 76(5), 650-661.

[22] Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence?. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences*, 42(4), 497-507.

[23] Turner R.M, Bird S.M., Higgins J.P.T. (2013) The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PloS ONE 2013; 8:e59202. doi:10.1371/journal.pone.0059202.

[24] Viechtbauer, W. (2010). Conducting meta-analyses in R with the metaphor package. *Journal of Statistical Software*, 36(3), 1-48

[25] Walker, E., Hernandez, A. V., Kattan, M.W. (2008). Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine*, 75(6), 431-439.

[26] Wang, L.L. (2010). Retrospective statistical power: Fallacies and recommendations. *Newborn and Infant Nursing Reviews*, 10(1), 55-59.

[27] Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass*, 2(6), 981-1022.

[28] Zumbo, B.D. and Hubley, A.M. (1998), A Note on Misconceptions Concerning Prospective and Retrospective Power. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47: 385-388. doi:10.1111/1467-9884.00139